Prédiction de la polysémie pour un terme biomédical

Juan Antonio Lossio-Ventura[1] — Clement Jonquet[1] — Mathieu Roche [1,2] — Maguelonne Teisseire [1,2]

[1] *LIRMM, Université de Montpellier, CNRS, Montpellier - France*
[2] *TETIS, Irstea, Cirad, Montpellier - France*

RÉSUMÉ. *La polysémie est la caractéristique d'un terme à avoir plusieurs significations. La prédiction de la polysémie est une première étape pour l'Induction de Sens (IS), qui permet de trouver des significations différentes pour un terme, ainsi que pour les systèmes d'extraction d'information. En outre, la détection de la polysémie est importante pour la construction et l'enrichissement de terminologies et d'ontologies. Dans cet article, nous présentons une nouvelle approche pour prédire si un terme biomédical est polysémique ou non, avec l'objectif à long terme d'enrichir les ontologies biomédicales après avoir désambiguïser les termes candidats. Cette approche est basée sur l'utilisation de techniques de méta-apprentissage, plus précisément sur des méta-descripteurs. Dans ce contexte, nous proposons la définition de nouveaux méta-descripteurs, extraits directement du texte, et d'un graphe de co-occurrences des termes. Notre méthode donne des résultats très satisfaisants, avec une exactitude et F-mesure de 0.978.*

ABSTRACT. *Polysemy is the capacity for a term to have multiple meanings. Polysemy prediction is a first step for Word Sense Induction (WSI), which allows to find different meanings for a term, as well as for Information Extraction (IE) systems. In addition, the polysemy detection is important for building and enriching terminologies and ontologies. In this paper, we present a novel approach to detect if a biomedical term is polysemic or not, with the long term goal of enriching biomedical ontologies after disambiguation of candidate terms. This approach is based on meta-learning techniques, more precisely on meta-features. We propose the definition of novel meta-features, extracted directly from the text dataset, as well as from a graph of coccurrent terms. Our method obtains very good results, with an Accuracy and F-mesure of 0.978.*

MOTS-CLÉS : *Polysémie, Induction de sens, Désambiguïsation, Méta-apprentissage, Méta-descripteurs, Apprentissage automatique, Terminologie biomédical, Graphes.*

KEYWORDS: *Polysemy, Word Sense Induction, Disambiguation, Meta-learning, Meta-features, Machine Learning, Biomedical Terminology, Graphs.*

1. Introduction

Le Web est de loin la plus grande source d'information disponible, qui évolue tous les jours avec des nouveaux contenus. Cette ressource contient des informations riches qui peuvent être liées à plusieurs domaines. La prise de décision, dans certaines organisations, dépend directement de la qualité de l'information existante sur le Web. C'est le cas par exemple en biomédecine, qui véhicule des connaissances à travers des publications très nombreuses (El-Rab *et al.*, 2013). Dans ce contexte, il existe de nombreuses méthodes pour extraire les informations pertinentes. La recherche sur le Web repose souvent sur des méthodes traditionnelles, Cependant, récemment ce problème a été traité par la recherche de concepts, qui analyse du texte pour extraire des instances de concepts associées aux requêtes des utilisateurs. Les ontologies sont très utiles pour l'identification de concepts. L'objectif principal est la création de nouvelle connaissance d'un domaine. Elles doivent être régulièrement enrichies par l'introduction de nouveaux termes.

Pour enrichir ces ontologies ou vocabulaires, il est nécessaire de connaître les sens possibles d'un terme, ce terme est alors candidat pour l'enrichissement ; c'est ce qu'on appelle l'Induction de Sens (IS). Une étape préliminaire à l'IS consiste à détecter si un terme est polysémique, puis si nécessaire, à faire une recherche exhaustive de ses sens. La détection de la polysémie est également très utile pour les systèmes de Recherche d'Information (RI), permettant d'avoir des meilleurs résultats lors de l'exécution des requêtes. Afin d'identifier les termes polysémiques, nous nous appuyons sur le principe du méta-apprentissage.

L'extraction de méta-descripteurs est la première étape du méta-apprentissage, qui exploite les méta-connaissances pour sélectionner la meilleure méthode pour la tâche de classification. Des techniques générales sont utilisées pour extraire des méta-descripteurs, et dans un objectif de classification, ces techniques peuvent être enrichies pour extraire encore plus de méta-descripteurs. Le méta-apprentissage a été appliqué dans des domaines différents mais, à notre connaissance, jamais pour la détection de la polysémie. Par conséquent, afin de relever le défi énoncé, nous profitons des avantages des méta-descripteurs. Dans ce contexte, nous proposons une nouvelle approche pour détecter si un terme est polysémique en définissant de nouveaux méta-descripteurs, extraits directement à partir de l'ensemble de données textuelles et d'un graphe induit à partir des co-occurrences des termes. Ces méta-descripteurs utilisent deux thésaurus de deux domaines différents (i.e., biomédical et agronomie), permettant d'identifier si un même terme est utilisé dans différents domaines.

À notre connaissance, une approche graphe n'a jamais été adoptée pour définir des méta-descripteurs. Dans ce travail, l'idée principale est de capturer les caractéristiques des données grâce à la forme structurelle et à la taille du graphe induit à partir de l'ensemble des données. Nous pourrons constater que cette approche obtient d'excellents résultats, surpassant les méthodes de la littérature avec 97.8% d'exactitude et de F-mesure.

L'article est organisé comme suit. Nous discutons d'abord des travaux connexes dans la Section 2. Ensuite, le modèle de classification et la construction de l'ensemble des données non polysémiques sont détaillés dans la Section 3. Les résultats sont présentés dans la Section 4, suivie par la conclusion et les perspectives dans la Section 5.

2. Travaux connexes

Nous nous sommes intéressés à la détection de la polysémie afin d'enrichir les terminologies ou ontologies en utilisant des approches d'Induction de Sens. Auparavant nous proposons d'identifier le cas où un terme est polysémique en utilisant des méthodes fondées sur le méta-apprentissage. Un aperçu des méthodes de méta-apprentissage et de détection de polysémie est proposé ci-après.

2.1. *Détection de la polysémie*

Une tâche liée à la détection de la polysémie concerne la détection d'ambiguïté d'un terme (term ambiguity detection - TAD) lié à un domaine (Baldwin *et al.*, 2013). Par exemple, étant donné un terme tel que *Brave* et une catégorie comme *film*, la tâche est de donner une réponse binaire pour savoir si toutes les instances de *Brave* référencent le film du même nom. Ceci est similaire à des problèmes bien étudiés de désambiguïsation des entités nommées (named entity disambiguation - NED) et désambiguïsation linguistique (word sense disambiguation - WSD). Ces tâches supposent que le nombre de sens d'un terme est spécifié. Cela rend ces tâches inadaptées pour l'enrichissement de terminologies.

Une tâche qui nécessite de la détection de la polysémie est l'Induction de Sens, qui permet de calculer à la fois le nombre de sens d'un terme, et ce qu'ils représentent. L'IS utilise des techniques non supervisées pour identifier automatiquement l'ensemble des sens d'un terme. Les principales approches d'IS proposées sont classées en quatre types (Navigli, 2012) : i) *Clustering du contexte* : le profil de distribution des mots exprime implicitement leur sémantique. Une approche bien connue est l'algorithme de la discrimination du contexte des groupes (Schütze, 1998 ; Van de Cruys et Apidianaki, 2011) ; ii) *Clustering de mots* : regroupe les mots qui sont sémantiquement similaires afin de découvrir un sens, par exemple nous pouvons citer le travail de (Pantel et Lin, 2002) ; iii) *Graphes de co-occurrences* : ces techniques ont le même principe que les approches de clustering de mots, mais elles utilisent des graphes de co-occurrences de mots pour identifier l'ensemble des sens d'un mot. Plusieurs travaux ont été développés sur la base de ces techniques (Navigli et Crisafulli, 2010), ou des algorithmes tels que HyperLex (Véronis, 2004), Pagerank (Agirre *et al.*, 2006 ; Agirre et Soroa, 2009) ; et iv) *Clustering probabiliste* : l'objectif est de formaliser l'IS dans un modèle génératif. Pour chaque mot ambigu une distribution de sens est tracée. Puis, le contexte de mots est généré selon cette distribution, e.g., (Lau *et al.*, 2012 ; Brody et Lapata, 2009).

2.2. *Méta-apprentissage*

Le méta-apprentissage est l'étude des méthodes qui exploitent des méta-descripteurs pour obtenir des modèles et des solutions efficaces pour le processus d'apprentissage automatique (Bhatt *et al.*, 2013). En effet, son but final est de construire, de manière automatique, des modèles adaptés pour un corpus donné (Duch *et al.*, 2011). Il y a deux tâches principales en méta-apprentissage. La première est la caractérisation du corpus avec les méta-descripteurs qui constituent les méta-données pour le méta-apprentissage. La deuxième est l'apprentissage au niveau méta, qui développe la méta-connaissance pour sélectionner l'algorithme approprié pour la classification (Peng *et al.*, 2002). Deux stratégies principales ont été développées afin de caractériser un corpus pour suggérer l'algorithme le plus approprié pour un corpus spécifique (Peng *et al.*, 2002). La première décrit les propriétés du corpus à l'aide de mesures statistiques et d'information. La seconde, considère un corpus caractérisé par la performance (e.g., exactitude) d'un ensemble de classificateurs, appelée landmarking. Dans ce travail, nous nous concentrerons sur la première stratégie.

Les mesures les plus fréquentes pour caractériser des corpus, sont la fréquence, la moyenne, l'écart-type, etc. Ces mesures ont été utilisées comme base des méta-descripteurs. Par exemple, (Peng *et al.*, 2002) ont extrait 15 méta-descripteurs de trois types : généraux, statistiques, et fondés sur la théorie de l'information. Des caractéristiques supplémentaires ont été proposées, dérivées des transformations de celles existantes, et des lignes directrices ont été fixées pour sélectionner les plus informatives. Par exemple, dans (Castiello *et al.*, 2005), 9 nouveaux méta-descripteurs ont été proposés comprenant les trois types mentionnés ci-dessus. D'autres méta-descripteurs statistiques ont été présentés dans d'autres travaux. Par exemple, (Reif *et al.*, 2012b) présentent une nouvelle approche pour la construction de méta-descripteurs plus informatifs en utilisant une méthode à deux étapes basée sur des méta-descripteurs traditionnels. Les méta-descripteurs proposés sont en mesure de décrire les différences des corpus qui ne sont pas accessibles à l'aide des méta-mesures. En outre, ils ont ajouté une méthode de sélection de caractéristique supplémentaire afin de retenir, de manière automatique, les mesures les plus utiles. Un travail similaire de (Reif *et al.*, 2012a) s'appuie sur une nouvelle fonction de génération de données pour créer des corpus ayant des caractéristiques spécifiques. Ceci peut être utile pour le développement et l'évaluation des systèmes de méta-apprentissage.

À notre connaissance, dans la littérature, il n'existe pas d'approche d'extraction de méta-descripteurs fondés sur un graphe induit à partir des co-occurrences des termes. Le graphe est une structure très utile grâce à ses propriétés spécifiques. Les graphes permettent d'obtenir des résultats très satisfaisants pour ce type de tâches comme nous allons le décrire dans la section suivante.

3. Vers la prédiction de la polysémie basée sur les nouveaux méta-descripteurs

Dans cette section, nous présentons la méthodologie proposée pour déterminer si un terme biomédical est polysémique ou non. Tout d'abord, nous présentons les principaux méta-descripteurs qui servent à caractériser le corpus. Pour créer ces nouveaux méta-descripteurs, nous appliquons des mesures statistiques et nous utilisons les thésaurus UMLS[1] et AGROVOC[2], respectivement un thésaurus biomédical (6 000 000 termes en anglais) et alimentaire/agronomique (40 000 termes en anglais).

Notre intuition concernant l'utilisation de deux thésaurus différents mais néanmoins connexes, est de déterminer si un terme apparaît dans les contextes propres aux thésaurus. Dans ce cas, nous pourrons supposer qu'un tel terme est polysémique. La Figure 1 synthétise le processus illustrant notre approche qui est détaillée ci-après.

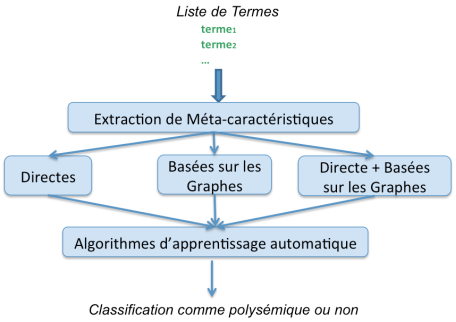

Figure 1. *Méthodologie pour la prédiction de la polysémie.*

3.1. *Méta-descripteurs*

Nous présentons de nouveaux méta-descripteurs basés sur des mesures statistiques pour caractériser notre corpus biomédical. Ils sont extraits directement à partir du corpus et à partir d'un graphe induit. Nous sélectionnons des algorithmes d'apprentissage appropriés pour déterminer si un terme est polysémique. L'idée principale, comme mentionné précédemment, est de définir les descripteurs du corpus grâce à la forme et à la taille du graphe induit à partir du corpus (Lossio-Ventura *et al.*, 2014c). Au total 23 méta-descripteurs sont ainsi proposées, 11 directes (c.f. Section 3.1.1) et 12 à partir du graphe induit (c.f. Section 3.1.2). Leur efficacité est illustrée en comparant les résultats obtenus par l'application des différents algorithmes d'apprentissage supervisé.

Notation : pour chaque terme t il existe un ensemble A_t des titres/résumés extraits de PubMed[3], $a \in A_t$, est un titre/résumé associé à un unique sens ou plusieurs sens pour les termes polysémiques.

1. `http://www.nlm.nih.gov/research/umls/`
2. `http://aims.fao.org/agrovoc`
3. Base de données des publications médicales, `www.ncbi.nlm.nih.gov/pubmed`

3.1.1. *Méta-descripteurs directs*

Les premiers méta-descripteurs de notre proposition sont extraits directement du corpus, sans la construction d'un modèle (e.g., un arbre, un graphe). Ils s'appuient sur des mesures statistiques qui reposent principalement sur le comptage des termes présents dans UMLS et AGROVOC, les deux thésaurus utilisés. Ils sont listés dans le Tableau 1.

1) Nombre de mots : représenté par $nWords(t)$, ceci correspond au nombre de mots qui composent le terme t. Par exemple *nWords(Lung cancer)*$= 2$.
2) Nombre de termes UMLS : représenté par $termsU(t)$, ceci correspond au nombre de termes UMLS trouvés dans l'ensemble de titres/résumés A_t de t.
3) Minimum de termes UMLS : dénoté par $minU(t)$, représente le nombre minimum de termes UMLS contenus dans chaque a de A_t. $$minU(t) = min(termsU(a_1), termsU(a_2), ...)$$
4) Maximum de termes UMLS : dénoté par $maxU(t)$, représente le nombre maximum de termes UMLS trouvés dans chaque a de A_t. $$maxU(t) = max(termsU(a_1), termsU(a_2), ...)$$
5) Moyenne de termes UMLS : dénoté par $meanU(t)$, représente la moyenne des nombres de termes UMLS trouvés pour chaque a de A_t. $$meanU(t) = \frac{1}{n} \times \sum_{i=1}^{n} termsU(a_i)$$
6) Écart type de termes UMLS : dénoté par $sdU(t)$, représente l'écart type des nombres de termes UMLS trouvés pour chaque a de A_t. $$sdU(t) = \frac{1}{n-1} \times \sqrt{\sum_{i=1}^{n} (termsU(a_i) - meanU(t))^2}$$
7) Nombre de termes AGROVOC : dénoté par $termsA(t)$, ceci correspond au nombre de termes AGROVOC trouvés dans l'ensemble de titres/résumés A_t de t.
8) Minimum de termes AGROVOC : dénoté par $minA(t)$, représente le nombre minimum de termes AGROVOC contenus dans chaque a de A_t. $$minA(t) = min(termsA(a_1), termsA(a_2), ...)$$
9) Maximum de termes AGROVOC : dénoté par $maxU(t)$, représente le nombre maximum de termes AGROVOC trouvés dans chaque a de A_t. $$maxA(t) = max(termsA(a_1), termsA(a_2), ...)$$
10) Moyenne de termes AGROVOC : dénoté par $meanU(t)$, représente la moyenne des nombres de termes AGROVOC trouvés pour chaque a de A_t. $$meanA(t) = \frac{1}{n} \times \sum_{i=1}^{n} termsA(a_i)$$
11) Écart type de termes AGROVOC : dénoté par $sdU(t)$, représente l'écart type des nombres de termes AGROVOC trouvés pour chaque a de A_t. $$sdA(t) = \frac{1}{n-1} \times \sqrt{\sum_{i=1}^{n} (termsA(a_i) - meanA(t))^2}$$

Tableau 1. *Méta-descripteurs directs.*

3.1.2. *Méta-descripteurs basés sur les graphes*

Comme mentionné précédemment, nous avons choisi d'utiliser la structure du graphe pour caractériser notre corpus. De cette façon, nous pouvons exploiter les propriétés du graphe, tels que les voisins, les poids des arêtes, la taille. Ainsi, nous avons construit un graphe pour chaque terme t et chaque graphe est indépendant des autres.

Construction du graphe : un graphe (par exemple, Figure 2) est construit pour chaque terme biomédical. Les sommets représentent les termes, et les arêtes (non orientées) désignent les relations de co-occurrences entre les termes. Les co-occurrences entre les termes sont mesurées comme le degré de relation entre deux termes dans un corpus initial. Cette relation est statistique et relie tous les termes co-occurrents sans tenir compte de leur signification ou de leur fonction dans le texte. Chaque graphe est construit avec les premiers 1000 premiers termes extraits avec l'application BIOTEX à partir de A_t. BIOTEX (Lossio-Ventura *et al.*, 2014a) est une application d'extraction automatique de termes biomédicaux qui met à disposition un ensemble de mesures telles que C-value, *TFIDF*, ou des combinaisons de ces mesures (Lossio-Ventura *et al.*, 2014b). Dans nos travaux, nous utilisons le *coefficient de Dice*, une mesure bien connue dans la littérature, pour calculer le degré de co-occurrence entre deux termes x et y, défini par la formule suivante :

$$D(x, y) = \frac{2 \times |x \cap y|}{|x| + |y|} \quad [1]$$

Où $|x|$ et $|y|$ représentent le nombre de titres/résumés où nous trouvons x et y, respectivement ; $|x \cap y|$ est le nombre de titres/résumés partagés par les deux termes ; $D(x, y)$ est la mesure de similarité entre x et y et varie entre 0 et 1.

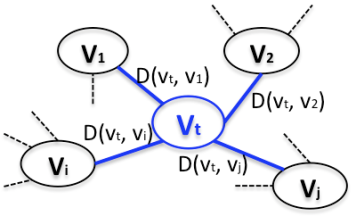

Figure 2. *Graphe créé pour le terme t.*

Dans la Figure 2, v_t représente le sommet pour le terme t, v_i représente un sommet i dans le graphe pour un terme i voisin de t. $\mathrm{N}(v_i)$ est le voisinage de v_i, $|\mathrm{N}(v_i)|$ le nombre de voisins de v_i, r_j le voisin j de v_i, $weight(v, r_j)$ le poids de l'arête entre v_i et son voisin r_j, ceci signifie que $weight(v_i, r_j) = D(v_i, r_j)$.

Le Tableau 2 liste l'ensemble des méta-descripteurs basés sur la structure du graphe. La taille de chaque graphe est de 1 000 sommets et les possibles relations de co-occurrences entre les sommets.

1) Nombre de voisins : ceci correspond au nombre de voisins du sommet v_t dans le graphe créé pour t. $$ng(v_t) = |\mathrm{N}(v_t)|$$
2) Somme de poids d'arêtes : dénoté par *sum*, représente la somme des poids d'arêtes spécifiquement pour le sommet v_t dans le graphe créé pour t. $$sum(v_t) = \sum_{j=1}^{ng(v_t)} weight(v_t, r_j)$$
3) Minimum du nombre de voisins : dénoté par *minNG*, représente le nombre minimum de voisins pour chaque v_i du graphe créé pour t. $$minNG(t) = min(ng(v_1), ng(v_2), ...)$$
4) Maximum du nombre de voisins : dénoté par *maxNG*, représente le nombre maximum de voisins pour chaque v_i du graphe créé pour t. $$maxNG(t) = max(ng(v_1), ng(v_2), ...)$$
5) Moyenne du nombre de voisins : dénoté par *meanNG*, représente la moyenne des nombres de voisins de chaque v_i dans le graphe créé pour t. $$meanNG(t) = \frac{\sum_{i=1}^{1000} ng(v_i)}{1000}$$
6) Écart type du nombre de voisins : dénoté par *sdNG*, représente l'écart type des nombres de voisins de chaque v_i dans le graphe créé pour t. $$sdNG(t) = \frac{\sqrt{\sum_{i=1}^{1000} (ng(v_i) - meanNG(t))^2}}{1000-1}$$
7) Somme minimum des poids d'arêtes : dénoté par *minSUM*, représente la somme minimum des poids d'arêtes pour chaque v_i du graphe créé pour t. $$minSUM(t) = min(sum(v_1), sum(v_2), ...)$$
8) Somme maximum des poids d'arêtes : dénoté par *maxSUM*, représente la somme maximum des poids d'arêtes pour chaque v_i du graphe créé pour t. $$maxSUM(t) = max(sum(v_1), sum(v_2), ...)$$
9) Moyenne de la somme des poids d'arêtes : dénoté par *meanSUM*, représente la moyenne de des somme des poids d'arêtes pour chaque v_i du graphe créé pour t. $$meanSUM(t) = \frac{\sum_{i=1}^{1000} sum(v_i)}{1000}$$
10) Écart type de la somme des poids d'arêtes : dénoté par *sdSUM*, représente l'écart type de des somme des poids d'arêtes pour chaque v_i du graphe créé pour t. $$sdSUM(t) = \frac{\sqrt{\sum_{i=1}^{1000} (sum(v_i) - meanSUM(t))^2}}{1000-1}$$
11) Nombre de voisins UMLS : représenté par *ngUMLS*, ceci correspond au nombre de voisins du sommet v_t, du graphe créé pour t, qui appartiennent à UMLS. $$ngUMLS(v_t) = |\mathrm{N}(v_t)|_{r_j \in UMLS}$$
12) Somme des poids d'arêtes UMLS : représenté par *sumUMLS*, représente la somme des poids d'arêtes de v_t, qui sont dans UMLS, dans le graphe créé pour t. $$sumUMLS(v_t) = \sum_{j=1}^{ngUMLS(v_t)} weight(v_t, r_j)$$

Tableau 2. *Méta-descripteurs extraits à partir du graphe induit.*

3.1.3. *Exemple*

Nous montrons avec un exemple pratique comment calculer les mesures pour caractériser notre corpus. Le Tableau 3 montre un ensemble de trois titres/résumés pour le terme *yellow fever*, ce fragment représente un extrait de notre corpus. Les termes trouvés dans UMLS sont en bleu et les termes dans AGROVOC sont encadrés en rouge. Notez que le terme *yellow fever* se trouve à la fois dans UMLS et AGROVOC.

Id	Titres/Résumés
a_1	"***Risks*** of ***travel***, benefits of a ***specialist*** consult. If ***patients*** are planning to ***travel*** to developing countries, their ***primary care physicians*** can counsel them on various medical ***risks***, especially traveler's ***diarrhea***, and offer to update their ***immunizations***. However, travelers to areas where there is a ***risk*** of ***malaria***, <t> ***yellow fever*** </t>, or other tropical ***diseases*** should be referred to a ***specialist***."
a_2	"***Herpes zoster*** after <t> ***yellow fever*** </t> ***vaccination***. An immunocompetent 64-year-old ***women*** presented with brachial ***herpes zoster*** (HZ) ***infection*** 3 days after ***vaccination*** against ***yellow fever*** (YF). The lesions disappeared after antiviral ***treatment***. There are very few ***reports*** of a possible ***association*** between YF ***vaccination*** and HZ ***infection***. This case supports the importance of continuing ***surveillance*** of vaccine adverse events."
a_3	"Broadening the horizons for <t> ***yellow fever*** </t> : new uses for an old vaccine. The vaccine against ***yellow fever*** is one of the safest and most effective ever developed. With an outstanding record in ***humans***, has this live attenuated vaccine been overlooked as a promising vector for the ***development*** of ***vaccines*** against pathogens outside its own genus ? Recent studies, including a ***report*** by Tao et al. on page 201 of this issue, have sparked renewed interest."

Tableau 3. *Extrait de titres et résumés associés au terme* ***"Yellow Fever"***

Les ***Termes trouvés dans UMLS (27)*** sont : association, development, diarrhea, diseases, fever, herpes zoster, humans, immunizations, infection, malaria, patients, physicians, primary care, primary care physicians, report, reports, risk, risks, specialist, surveillance, travel, treatment, vaccination, vaccines, women, yellow fever, zoster.

Les Termes trouvés dans AGROVOC (22) sont : countries, developing countries, development, diseases, events, fever, genus, humans, infection, lesions, malaria, pathogens, patients, physicians, planning, reports, risk, uses, vaccination, vaccines, women, yellow fever.

La Figure 3 montre un sous-graphe créé avec tous les titres/résumés pour le terme *yellow fever*. Ce graphe contient les termes UMLS encerclés et écrits en bleu. Le Tableau 4 présente les valeurs de tous les méta-descripteurs.

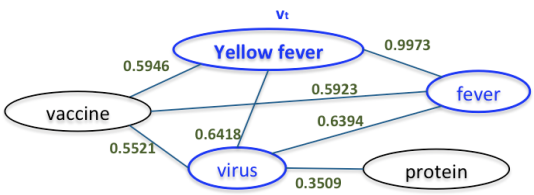

Figure 3. *Sous-graphe créé pour le terme* ***Yellow Fever***

Extraction de méta-descripteurs directs		
Item	Mesure	Commentaire
1	$nWords(t) = 2$	"yellow fever" contient deux mots.
2	$termsU(t) = 27$	Le nombre total de différents termes UMLS trouvés dans l'ensemble de titres/résumés.
3	$minU(t) = 6$	Le nombre minimum de différents termes UMLS trouvés pour chaque titre/résumé a_i de t. Cela veut dire $min(termsU(a_1), termsU(a_2), termsU(a_3)) = min(13, 11, 6) = 6$
4	$maxU(t) = 13$	Le nombre maximum de différents termes UMLS trouvés pour chaque titre/résumé a_i de t. Cela veut dire $max(termsU(a_1), termsU(a_2), termsU(a_3)) = max(13, 11, 6) = 13$
5	$meanU(t) = 12$	$\frac{\sum_{i=1}^{n} termsU(a_i)}{n} = \frac{\sum_{i=1}^{3} termsU(a_i)}{3} = \frac{13+11+6}{3} = 10$
6	$sdU(t) = 2.55$	$\frac{\sqrt{\sum_{i=1}^{n}(termsU(a_i)-meanU(t))^2}}{n-1} = \frac{\sqrt{\sum_{i=1}^{3}(termsU(a_i)-10)^2}}{3-1} = \frac{\sqrt{(13-10)^2+(11-10)^2+(6-10)^2}}{2} = 2.55$
7	$termsA(t) = 22$	Le nombre total de termes AGROVOC trouvés dans l'ensemble de titres/résumés.
8	$minA(t) = 8$	Le nombre minimum de différents termes AGROVOC trouvés pour chaque titre/résumé a_i de t. Cela veut dire $min(termsA(a_1), termsA(a_2), termsA(a_3)) = min(10, 8, 8) = 8$
9	$maxA(t) = 10$	Le nombre maximum de différents termes AGROVOC trouvés pour chaque titre/résumé a_i de t. Cela veut dire $max(termsA(a_1), termsA(a_2), termsA(a_3)) = max(10, 8, 8) = 10$
10	$meanA(t) = 8.67$	$\frac{\sum_{i=1}^{n} termsA(a_i)}{n} = \frac{\sum_{i=1}^{3} termsA(a_i)}{3} = \frac{10+8+8}{3} = 8.67$
11	$sdA(t) = 0.82$	$\frac{\sqrt{\sum_{i=1}^{n}(termsA(a_i)-meanA(t))^2}}{n-1} = \frac{\sqrt{\sum_{i=1}^{3}(termsA(a_i)-8.67)^2}}{3-1} = \frac{\sqrt{(10-8.67)^2+(8-8.67)^2+(8-8.67)^2}}{2} = 0.82$
Extraction de méta-descripteurs basés sur les graphes		
Item	Mesure	Commentaire
1	$ng(v_t) = 3$	le sommet v_t "yellow fever" a 3 voisins.
2	$sum(v_t) = 2.2337$	$\sum_{j=1}^{ng(v_t)} weight(v_t, r_j) = \sum_{j=1}^{3} weight(v_t, r_j) = weight(\text{yellow fever, vaccine}) + weight(\text{yellow fever, fever}) + weight(\text{yellow fever, virus}) = 0.5946 + 0.9973 + 0.6418 = 2.2337$
3	$minNG(t) = 1$	Le nombre minimum de voisins de chaque v_i du graphe créé pour t. Cela veut dire $min(ng(v_t), ng(v_1), ng(v_2), ng(v_3), ng(v_4)) =$ $min(ng(\text{yellow fever}), ng(\text{vaccine}), ng(\text{fever}), ng(\text{virus}), ng(\text{protein})) = min(3, 3, 3, 4, 1) = 1$
4	$maxNG(t) = 4$	Le nombre maximum de voisins de chaque v_i du graphe créé pour t. Cela veut dire $max(ng(v_t), ng(v_1), ng(v_2), ng(v_3), ng(v_4)) =$ $max(ng(\text{yellow fever}), ng(\text{vaccine}), ng(\text{fever}), ng(\text{virus}), ng(\text{protein})) = max(3, 3, 3, 4, 1) = 4$
5	$meanNG(t) = 2.8$	$\frac{\sum_{i=1}^{5} ng(v_i)}{5} = \frac{\sum_{i=1}^{5} ng(v_i)}{5} = \frac{3+3+3+4+1}{5} = 2.8$ // dans ce cas le dénominateur a la valeur de 5 au lieu de 1000, parce que le sous-graphe ne que contient que 5 sommets.
6	$sdNG(t) = 0.55$	$\frac{\sqrt{\sum_{i=1}^{5}(ng(v_i)-meanNG(t))^2}}{5-1} = \frac{\sqrt{\sum_{i=1}^{5}(ng(v_i)-2.8)^2}}{5-1} = \frac{\sqrt{(3-2.8)^2+(3-2.8)^2+(3-2.8)^2+(4-2.8)^2+(1-2.8)^2}}{4} = 0.55$ // dans ce cas le dénominateur a la valeur de 5 au lieu de 1000, parce que le sous-graphe ne que contient que 5 sommets.
7	$minSUM(t) = 0.3509$	La somme minimum des poids d'arêtes pour chaque v_i dans le graphe créé pour t. Cela veut dire $min(sum(v_t), sum(v_1), sum(v_2), sum(v_3), sum(v_4)) =$ $min(sum(\text{yellow fever}), sum(\text{vaccine}), sum(\text{fever}), sum(\text{virus}), sum(\text{protein})) =$ $min(2.2337, 1.739, 2.229, 2.1842, 0.3509) = 0.3509$
8	$maxSUM(t) = 2.2337$	La somme maximum des poids d'arêtes pour chaque v_i dans le graphe créé pour t. Cela veut dire $max(sum(v_t), sum(v_1), sum(v_2), sum(v_3), sum(v_4)) =$ $max(sum(\text{yellow fever}), sum(\text{vaccine}), sum(\text{fever}), sum(\text{virus}), sum(\text{protein})) =$ $max(2.2337, 1.739, 2.229, 2.1842, 0.3509) = 2.2337$
9	$meanSUM(t) = 1.7474$	$\frac{\sum_{i=1}^{5} sum(v_i)}{5} = \frac{\sum_{i=1}^{5} sum(v_i)}{5} = \frac{2.2337+1.739+2.229+2.1842+0.3509}{5} = 1.7474$ //dans ce cas le dénominateur a la valeur de 5 au lieu de 1000, parce que le sous-graphe ne que contient que 5 sommets.
10	$sdSUM(t) = 0.40$	$\frac{\sqrt{\sum_{i=1}^{5}(sum(v_i)-meanSUM(t))^2}}{5-1} = \frac{\sqrt{\sum_{i=1}^{5}(sum(v_i)-1.7474)^2}}{5-1} = 0.40$
11	$ngUMLS(v_t) = 2$	Le sommet v_t "yellow fever" a deux voisins UMLS (*fever* et *virus*).
12	$sumUMLS(v_t) = 1.6391$	$\sum_{j=1}^{ngUMLS(v_t)} weight(v_t, r_j) = \sum_{j=1}^{2} weight(v_t, r_j) =$ $weight(\text{yellow fever, fever}) + weight(\text{yellow fever, virus}) = 0.9973 + 0.6418 = 1.6391$

Tableau 4. *Extraction de méta-descripteurs pour le terme t =* ***Yellow Fever***

3.2. *Algorithmes d'apprentissage supervisé*

Dans la suite, nous allons utiliser les méta-descripteurs décrits précédemment comme entrées de divers algorithmes d'apprentissage supervisé de la littérature. Nous avons utilisé pour cela le logiciel Weka[4] (Hall *et al.*, 2009) (avec un paramétrage des valeurs par défaut pour chaque algorithme). Les algorithmes que nous avons comparés sont :

Naives Bayes (NB)	AdaBoost (AB)
Tree Decision (TD)	SVM (SVM)
Meta Bagging (MB)	M5P Tree (M5P)
Multilayer Perceptron (NN)	MultiClassClassifier Logistic (MCC)

4. Données et résultats

Pour évaluer notre approche, nous avons extrait les méta-descripteurs à partir d'un *Gold Standard Corpus* composé de l'union d'un corpus polysémique préexistant et de la construction d'un corpus non polysémique que nous avons construit. Tous les algorithmes de classification cités précédemment ont été comparés dans le but de produire la meilleure prédiction relative à la polysémie ou non d'un terme biomédical. Nous expliquons tous les résultats obtenus avec ces algorithmes dans la Section 4.2.

4.1. *Construction du Gold Standard Corpus*

4.1.1. *Corpus polysémique*

Nous utilisons le corpus MSH WSD[5] (Jimeno-Yepes *et al.*, 2011), qui est un benchmark reconnu pour la désambiguïsation dans le domaine biomédical. Ce corpus est composé de 106 abréviations ambiguës, 88 termes ambigus et 9 éléments qui sont une combinaison des abréviations et des termes ambigus. Cela donne un total de 203 entités ambiguës. Pour chaque terme/abréviation ambigu, le corpus contient un maximum de 100 instances par sens obtenu à partir de PubMed.

4.1.2. *Construction du corpus non polysémique*

De la même manière dont le corpus MSH WSD a été créé, nous avons construit un corpus non polysémique[6] à l'aide du metathésaurus UMLS et de PubMed. Dans le corpus MSH WSD, pour chaque terme il existe moins de 500 titres/résumés, et ce nombre est différent pour chaque terme. Par conséquent, nous avons besoin de construire un corpus non polysémique diminuant le biais produit du fait de la variabilité des données. Ce biais affecte les données et la tâche d'apprentissage. Pour contourner ce pro-

4. `http://www.cs.waikato.ac.nz/ml/weka/`
5. `http://wsd.nlm.nih.gov/`
6. Contactez nous pour avoir accès à ce corpus

blème, nous créons le nouveau corpus comme un "miroir" du nombre de titres/résumés du corpus MSH WSD. Le Tableau 5 illustre ce principe. La première ligne montre le terme *Cold*, lequel dispose de 260 titres/résumés, donc pour le nouveau corpus, nous sélectionnons le terme $Terme_1$ et extrayons le même nombre de titres/résumés, c'est-à-dire 260.

Terme polysémique	**Nombre de Titres/Résumés**	**Terme non polysémique**	**Nombre de Titres/Résumés**
Cold	260	$Terme_1$	260
Cortical	297	$Terme_2$	297
Yellow Fever	183	$Terme_4$	183
...	...	...	...

Tableau 5. *Principe de miroir pour la construction du corpus non polysémique.*

Les étapes pour la création du corpus non polysémique sont :

1) Tout d'abord, nous sélectionnons tous les termes contenus dans MeSH (un des thésaurus inclus dans UMLS). Puis nous considérons l'ensemble d'UMLS pour identifier les termes non ambigus (ceux qui sont associés à un seul concept, UMLS/CUI). Nous filtrons cette liste, en ne prenant que ceux qui sont non polysémiques.

2) Les termes UMLS contiennent plusieurs signes (symboles), nous les nettoyons en éliminant tous les termes contenant (; , ? ! : { } [ ]).

3) Nous faisons une recherche sur PubMed afin de connaître le nombre d'articles par terme. La condition utilisée pour cette recherche est que le terme doit apparaître à la fois dans le titre et dans le résumé. Si le nombre d'articles retourné est supérieur à 500, alors nous prenons le terme et les données correspondantes.

4) Nous choisissons aléatoirement 203 termes non polysémiques et nous extrayons leur contenu à partir d'un fichier XML obtenu à partir d'une autre requête sur PubMed. À la fin de cette étape, nous avons constitué le corpus.

4.2. *Expérimentations et résultats*

Dans cette section, nous décrivons les expérimentations effectuées pour évaluer la performance des méta-descripteurs proposés (23 au total). Nous avons réalisé des expérimentations de notre approche avec une validation croisée de 10%. Les résultats sont évalués en termes d'*Exactitude (E), Précision (P), Rappel (R)* et *F-mesure (F)*. Dans la Section 4.2.1, des expérimentations sont effectuées avec seulement les méta-descripteurs directs, dans la Section 4.2.2, les méta-descripteurs basés sur les graphes sont évalués. Nous avons voulu également explorer la performance des méta-descripteurs en combinant les 11 méta-descripteurs directs avec les 12 basés sur les graphes, ces résultats sont présentés dans la Section 4.2.3. Une comparaison de notre approche avec les autres de la littérature demeure difficile à réaliser car, à notre connaissance, il n'existe pas de travaux focalisés dans la détection de la polysémie dont le résultat est binaire (i.e., vrai ou faux). La plupart des travaux concernant la polysémie consistent à identifier les sens corrects à un terme dans son contexte donné.

4.2.1. *Méta-descripteurs directs*

Le Tableau 6 montre les résultats obtenus sur notre corpus précédemment décrit uniquement avec les méta-descripteurs directs. Nous constatons que l'algorithme M5P Tree obtient les meilleurs résultats, avec une exactitude de *0.921*. Cela signifie que les algorithmes supervisés sur nos méta-descripteurs directs ont classé correctement 92% des instances (polysémiques ou non).

	E	*P*	*R*	*F*
NB	0.860	0.863	0.860	0.859
AB	0.897	0.903	0.897	0.896
TD	0.879	0.882	0.879	0.879
SVM	0.919	0.922	0.919	0.919
MB	0.892	0.896	0.892	0.891
M5P	**0.921**	**0.925**	**0.921**	**0.921**
NN	0.906	0.907	0.921	0.906
MCC	0.914	0.915	0.914	0.914

Tableau 6. *Résultats avec les 11 méta-descripteurs directs.*

4.2.2. *Méta-descripteurs basés sur les graphes*

Le Tableau 7 montre les résultats obtenus uniquement avec les méta-descripteurs basés sur les graphes. Nous constatons que l'algorithme Meta Bagging obtient les meilleurs résultats, avec une exactitude de 92.1%. Les résultats obtenus avec les algorithmes supervisés sont différents pour les deux types de méta-descripteurs. La raison principale est que les méta-descripteurs et leurs valeurs sont différents selon les deux approches utilisées.

	E	*P*	*R*	*F*
NB	0.860	0.863	0.860	0.859
AB	0.899	0.900	0.899	0.899
TD	0.882	0.884	0.882	0.882
SVM	0.874	0.875	0.874	0.874
MB	**0.921**	**0.922**	**0.921**	**0.921**
M5P	0.884	0.885	0.884	0.884
NN	0.906	0.907	0.906	0.906
MCC	0.914	0.914	0.914	0.914

Tableau 7. *Résultats avec les 12 méta-descripteurs basés sur les graphes.*

4.2.3. *Combinaison des deux types de méta-descripteurs*

Nous étudions l'effet de la combinaison des deux types de méta-descripteurs. Cette combinaison consiste à considérer les 23 méta-descripteurs des deux approches. Le Tableau 8 montre ces résultats. Nous constatons que le modèle de Réseau Neuronal - Multilayer Perceptron (NN) obtient d'excellents résultats, avec une exactitude de

97.8%. Ce tableau met aussi en relief que la performance la moins élevée (95.3% d'exactitude), reste supérieure aux résultats obtenus soit avec uniquement les méta-descripteurs directs soit avec uniquement les méta-descripteurs basés sur les graphes. Ces résultats confirment donc que la combinaison de deux types de méta-descripteurs est plus performante.

	E	*P*	*R*	*F*
NB	0.956	0.956	0.956	0.956
AB	0.975	0.976	0.975	0.975
TD	0.970	0.970	0.970	0.970
SVM	0.966	0.966	0.966	0.966
MB	0.970	0.970	0.970	0.970
M5P	0.963	0.963	0.963	0.963
NN	**0.978**	**0.978**	**0.978**	**0.978**
MCC	0.953	0.953	0.953	0.953

Tableau 8. *Combinaison des deux types de méta-descripteurs.*

5. Conclusions et perspectives

Dans cet article, nous avons présenté une nouvelle approche pour prédire si un terme est polysémique dans le domaine biomédical. La contribution est avant tout la définition de nouveaux méta-descripteurs, qui sont directement extraits du texte du corpus et basés sur un graphe de co-occurrence. Une deuxième contribution de ce travail, est la construction d'un corpus non polysémique.

Notre nouvelle approche est basée dans des techniques de méta-apprentissage, pour extraire les méta-descripteurs qui décrivent au mieux un corpus. Cela a permis une classification plus efficace (i.e., prédiction de la polysémie). Pour la classification, nous avons utilisé les algorithmes supervisés les plus connus avec l'ensemble de méta-descripteurs proposés. Ces méta-descripteurs sont extraits de deux façons. D'abord, ils sont extraits directement du corpus et correspondent aux caractéristiques les plus pertinentes et les plus rapides à obtenir. Ensuite, ils sont extraits d'un graphe, construit selon l'ensemble de données de chaque terme. Cela permet de profiter des avantages des propriétés des graphes afin de caractériser le corpus.

Les méta-descripteurs ont été évalués de trois manières différentes. Tout d'abord, nous avons évalué les méta-descripteurs directs. Ensuite, l'évaluation des méta-descripteurs basés sur les graphes. Et finalement, nous avons évalué la performance de la fusion de ces deux types de méta-descripteurs, obtenant les meilleurs résultats.

Différentes améliorations peuvent être prises en compte dans les travaux futurs. Par exemple, l'ajout de nouveaux méta-descripteurs sur la base d'autres thésaurus comme WORDNET constitue une perspective intéressante. Une autre perspective serait la réutilisation du graphe créé pour chaque terme polysémique, et d'appliquer certaines approches d'IS pour déterminer le nombre de sens que ce terme contient.

Notre approche vers l'enrichissement d'ontologies aidera les utilisateurs dans l'analyse, la création et la transformation d'information pertinente à partir de corpus volumineux. Nos futurs travaux pourront se concentrer sur la transformation du langage informel issu des données textuelles vers des représentations formelles. Nous nous intéressons également à comment combiner les méthodes traditionnelles de recherche d'information avec les portails Web (e.g., BioPortal[7]) pour apporter des améliorations mesurables aux utilisateurs. Ce travail peut aussi être utilisé pour détecter la polysémie d'un terme lors d'une requête.

Finalement, nous soulignons que les méta-descripteurs offrent des possibilités dans différents domaines grâce à des techniques générales qui peuvent être extrapolées à d'autres algorithmes pour les tâches d'extraction de descripteurs. Bien que l'utilisation de méta-apprentissage dans différents domaines a augmenté au cours des dernières années, il existe encore un grand nombre de contextes dans lesquels ces approches n'ont pas encore été étudiées.

Remerciements

Ce travail est financé par l'Agence Nationale de la Recherche de France sous le programme JCJC (ANR-12-JS02-0010), le projet SIFR[8], l'Université de Montpellier, le CNRS, le project IBC, et par le programme de bourses de FINCyT du Pérou.

6. Bibliographie

Agirre E., Martinez D., De Lacalle O. L., Soroa A., « Evaluating and optimizing the parameters of an unsupervised graph-based WSD algorithm », *Proceedings of the first workshop on graph based methods for natural language processing*, ACL, p. 89-96, 2006.

Agirre E., Soroa A., « Personalizing Pagerank for Word Sense Disambiguation », *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, ACL, p. 33-41, 2009.

Baldwin T., Li Y., Alexe B., Stanoi I. R., « Automatic Term Ambiguity Detection », *Proceedings of the 51st Annual Meeting of the ACL*, Sofia, Bulgaria, p. 804-809, 2013.

Bhatt N., Thakkar A., Ganatra A., Bhatt N., « Ranking of Classifiers based on Dataset Characteristics using Active Meta Learning », *International Journal of Computer Applications*, vol. 69, n° 20, p. 31-36, May, 2013.

Brody S., Lapata M., « Bayesian word sense induction », *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, ACL, p. 103-111, 2009.

Castiello C., Castellano G., Fanelli A. M., « Meta-data : Characterization of Input Features for Meta-Learning », *Modeling Decisions for Artificial Intelligence*, Springer, p. 457-468, 2005.

Duch W., Maszczyk T., Grochowski M., « Optimal Support Features for Meta-Learning », *Meta-Learning in Computational Intelligence*, Springer, p. 317-358, 2011.

7. http://bioportal.bioontology.org
8. https://sites.google.com/site/sifrproject/

El-Rab W. G., Zaiane O. R., El-Hajj M., « Biomedical text disambiguation using UMLS », *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ACM, p. 943-947, 2013.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., « The WEKA data mining software : an update », *ACM SIGKDD explorations newsletter*, vol. 11, n° 1, p. 10-18, 2009.

Jimeno-Yepes A. J., McInnes B. T., Aronson A. R., « Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation », *BMC bioinformatics*, vol. 12, n° 1, p. 223, 2011.

Lau J. H., Cook P., McCarthy D., Newman D., Baldwin T., « Word sense induction for novel sense detection », *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, ACL, p. 591-601, 2012.

Lossio-Ventura J. A., Jonquet C., Roche M., Teisseire M., « BIOTEX : A system for Biomedical Terminology Extraction, Ranking, and Validation », *ISWC'2014 : The 13th International Semantic Web Conference*, Riva del Garda, Italy, 2014a.

Lossio-Ventura J. A., Jonquet C., Roche M., Teisseire M., « Towards a mixed approach to extract biomedical terms from text corpus », *International journal of Knowledge Discovery in Bioinformatics*, vol. 4, n° 1, p. 1-15, 2014b.

Lossio-Ventura J. A., Jonquet C., Roche M., Teisseire M., « Yet Another Ranking Function for Automatic Multiword Term Extraction », *PolTAL'2014 : 9th International Conference on NLP*, n° 8686 in *LNAI*, Springer, Warsaw, Poland, p. 52-64, 2014c.

Navigli R., « A quick tour of word sense disambiguation, induction and related approaches », *SOFSEM 2012 : Theory and practice of computer science*, Springer, p. 115-129, 2012.

Navigli R., Crisafulli G., « Inducing Word Senses to Improve Web Search Result Clustering », *Proceedings of the 2010 conference on empirical methods in natural language processing*, ACL, p. 116-126, 2010.

Pantel P., Lin D., « Discovering Word Senses from Text », *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, p. 613-619, 2002.

Peng Y., Flach P. A., Soares C., Brazdil P., « Improved Dataset Characterisation for Meta-Learning », *Discovery Science*, Springer, p. 141-152, 2002.

Reif M., Shafait F., Dengel A., « Dataset generation for meta-learning », *35th German Conference on Artificial Intelligence*, 2012a.

Reif M., Shafait F., Dengel A., « Meta2-Features : Providing Meta-Learners More Information », *35th German Conference on Artificial Intelligence*, 2012b.

Schütze H., « Automatic word sense discrimination », *Computational linguistics*, vol. 24, n° 1, p. 97-123, 1998.

Van de Cruys T., Apidianaki M., « Latent Semantic Word Sense Induction and Disambiguation », *Proceedings of the 49th Annual Meeting of the ACL : Human Language Technologies-Volume 1*, ACL, p. 1476-1485, 2011.

Véronis J., « Hyperlex : lexical cartography for information retrieval », *Computer Speech & Language*, vol. 18, n° 3, p. 223-252, 2004.